DOCUMENT RESUME

ED 306 283 TM 013 142

AUTHOR Marso, Ronald N.; Pigge, Fred L.
TITLE The Status of Classroom Teachers' Test Construction Proficiencies: Assessments by Teachers, Principals, and Supervisors Validated by Analyses of Actual Teacher-Made Tests.
PUB DATE Mar 89
NOTE 39p.; Paper presented at the Annual Meeting of the National Council of Measurement in Education (San Francisco, CA, March 28-30, 1989).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Administrator Attitudes; Classroom Techniques; Elementary Secondary Education; Inservice Teacher Education; Preservice Teacher Education; *Principals; Public Schools; *Supervisors; Teacher Attitudes; Teacher Effectiveness; Teacher Evaluation; *Teacher Made Tests; *Test Construction; Test Use
IDENTIFIERS Ohio; *Teacher Competencies

ABSTRACT

The extent to which principals (n=313), supervisors (n=273), and teachers (n=313) agreed in their assessments of classroom teachers' testing proficiencies was studied in the Ohio public schools. There were 122 elementary and 191 secondary school teachers in the teacher sample. The validity of these perceptions was determined by a direct assessment of teachers' testing proficiencies or skills as demonstrated on teacher-made tests. Perceptual assessments of all three groups differed significantly. These assessments also differed from the demonstrated proficiencies of the teachers, evaluated by two judges from a sample of 175 tests made by these teachers, for a total of 6,529 test items and 455 item exercises. Lack of agreement was especially evident for teachers' test item writing and test format development skills and for the teachers' writing of items functioning at higher cognitive levels. Moderately high negative correlations were found between perceived and demonstrated proficiency assessments of teachers' test item writing skills, but moderate to high positive correlations were found between the three sets of perceptual assessments, with the highest correlation between assessments of teachers and supervisors. Teachers rated their testing proficiencies higher than did principals, and principals rated proficiencies higher than did supervisors. Results suggest that perceptions of educational staff may not be accurate guides for determining needs for inservice training. Both preservice and inservice teacher educators need to increase attention to test item writing skills and the writing of items that measure beyond simple knowledge levels. Seven tables present survey and assessment data. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED306283

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

RONALD N. MARSO
FRED L. PIGGE

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

The Status of Classroom Teachers' Test Construction Proficiencies: Assessments by Teachers, Principals, and Supervisors Validated by Analyses of Actual Teacher-Made Tests

Ronald N. Marso and Fred L. Pigge

College of Education and Allied Professions

Bowling Green State University

Bowling Green, Ohio 43403

A paper presented at the annual meeting of the

National Council on Measurement in Education

San Francisco, California

March 28-30, 1989

Running Head: TEST CONSTRUCTION PROFICIENCIES

TM013142

BEST COPY AVAILABLE

Abstract

This study was designed to ascertain the extent that principals (n=313), supervisors (n=273), and teachers (n=313) agreed in their assessments of classroom teachers' testing proficiencies and whether or not these perceptual evaluations would be validated by a direct assessment of teachers' testing proficiencies or skills as demonstrated on their teacher-made tests. It was found that principals', supervisors', and teachers' perceptual assessments differed or varied significantly from each other and also that these assessments differed from demonstrated teachers' testing proficiencies. Widespread lack of agreements among the four data sets were noted for the teachers' test item writing and test format development skills and also for the teachers' writing of items functioning at higher cognitive levels. Moderately high negative correlations (-.50s to -.70s) were found between the perceived and demonstrated proficiency assessments of teachers' test item writing skills, and moderate to high positive correlations (.40s to .90s) were found between the three sets of perceptual assessments with the highest correlation between teachers' and supervisors' assessments.

The Status of Classroom Teachers' Test Construction Proficiencies: Assessments by Teachers, Principals, and Supervisors Validated by Analyses of Actual Teacher-Made Tests

It has been stated that those of us in higher education have a limited understanding of the nature of teacher-made tests used in the K-12 classrooms of our nation (e.g. Dwyer, 1982; Fleming & Chambers, 1983; Stiggins & Bridgeford, 1985). Further, Gullickson (1984) suggested that we do not know whether teacher-made tests are used effectively or even how they are used in the classroom. Relatedly, some evidence exists which indicates that the content of teacher preservice tests and measurement courses designed to facilitate the effective development and use of teacher-made tests may not be meeting the needs of classroom teachers (e.g., Gullickson, 1986; Gullickson & Ellwein, 1985; Salmon-Cox, 1981).

A number of recent investigations have addressed questions related to the characteristics of teacher-made tests and to classroom teachers' testing attitudes and practices; however, the vast majority of these studies were limited to teacher self-report procedures or to second-hand (i.e., principal or supervisor) assessments of teachers' testing knowledge or practices. For example, Gullickson (1984) reported that classroom teachers felt that frequent classroom testing is

desirable and facilitates instruction; Gullickson and Ellwein (1985) found that very few teachers complete simple statistical analyses of their test scores such as the calculation of means and standard deviations; Rogers (1985) indicated that teachers believe that observations of student performance and product ratings are desirable supplements to paper and pencil tests; and Lambert (1980-81) identified a consensus related to the importance of teachers producing superior classroom tests among samples of state education legislative committee chairpersons, principal officials in state teacher associations, and deans of teacher education institutions.

Very few studies have been reported in the educational literature wherein direct analyses of teacher-made tests constituted the data for the study. In one such study Fleming and Chambers (1983) conducted an extensive analysis of 342 teacher-made tests and found that the test items comprising these tests functioned primarily at the knowledge level (averages over grade levels were from 69% to 94% but most items functioning at higher cognitive levels were found only on the math and science tests), that directions were absent on approximately one-third of the tests, that items frequently were not numbered consecutively throughout the tests, that grammatical, spelling, and punctuation type errors appeared on 15% to 20% of the tests, that a large portion of the tests were

handwritten and illegible, that the short response type items found on the tests tended to be ambiguous, and that one- or two-word stems were commonly found among the multiple-choice type items.

Two additional studies involving the direct analysis of samples of teacher-made tests were located in the literature; however, these studies were limited to just an assessment of the cognitive functioning levels of the items on science tests. Billeh (1974) examined 33 seventh through tenth grade science tests and concluded that 72% of the items found on the tests functioned at the knowledge level. Additionally, he noted that item cognitive functioning levels did not vary by grade level of the test or by the extent of the training of the teachers who had constructed the tests; that the biology tests contained more knowledge level items than did the physics or chemistry tests; and that the more experienced teachers when compared to the less experienced teachers used more knowledge level items on their tests. Similarly, Black (1980) reported a direct analysis of 48 secondary science tests and also found that item cognitive functioning levels did not vary by extent of teacher training but did vary by science subject area. Regarding the latter, this researcher found that knowledge level items constituted 94% of the biology, 66% of the chemistry, and 56% of the physics test items.

Purpose

The purpose of this investigation was to ascertain the extent that school building principals', teacher supervisors', and classroom teachers' perceptual assessments of teachers' testing proficiencies agreed and whether or not these perceptual evaluations would be validated by direct indepth assessments of actual tests made by the classroom teachers. Whether or not perceptual assessments accurately represent teachers' actual test construction skills as displayed on their teacher-made tests would appear to be a significant theoretical and practical research concern. Most of the currently available literature on teachers' classroom tests and testing practices and most of the data gathered for teacher inservice planning are heavily dependent upon perceptual assessments for their accuracy. A secondary purpose of this study was to determine whether teachers, principals, and supervisors perceive classroom teachers' testing proficiencies as being comparable to the level of teachers' other professional proficiencies, such as general classroom management, subject matter knowledge, etc.

The following three hypotheses were stated to provide direction for this investigation: One, principals, supervisors, and teachers will not differ significantly in their assessments of classroom teachers' test planning and test construction skills. Two, there will be high positive agreements between the

teachers', principals', and supervisors' perceived levels of teachers' testing skills, and in turn these perceptual assessments will be highly related to the levels of teachers' testing skills derived from direct analyses of the teachers' actual tests. And three, the supervisors, principals, and teachers will indicate that beginning teachers' proficiencies (or in the case of teachers, their present proficiencies) in classroom testing and evaluation skills are equivalent to or higher than their proficiencies in their other professional skills. Specifically, it is predicted that the supervisors, principals, and the teachers, themselves, will rate teachers' testing skills as being as high or higher than: a) knowledge of their subject areas, b) their other professional education skills, and c) their overall skills as educators.

Methods and Procedures

Subjects

The administrator subjects for the study consisted of random samples of approximately 800 principals and supervisors selected from the state directory of Ohio public schools with school system (city, exempted village, and county local), job assignment (principal and supervisor), and school grade level (elementary, middle, and secondary) classifications used as strata. A total of 586 (73%) usable survey responses were obtained after two follow-up contacts of nonrespondents. A

total of 229 supervisors, 313 building principals, and 44 individuals in related supervisory roles (coordinators of curriculum or instruction, etc.) returned usable and complete assessment forms.

The teacher subjects for the study consisted of approximately 600 teachers who had graduated from Bowling Green State University during the 1975-1986 period and who were teaching full-time in Ohio during the 1985-86 academic year. These individuals were identified by "matching" the social security numbers of BGSU graduates during this time period with a computer listing of social security numbers of all regular classroom teachers certified by the Ohio State Department of Education for the 1985-86 school year. Teachers with special certifications (e.g., art, music, special education, etc.) were excluded from the selection process. A return of usable responses from 313 (52%) regular classroom teachers after two follow-up contacts of nonrespondents was realized. The teachers responding to the survey instrument consisted of 122 elementary and 191 secondary teachers.

Test Sample and Analysis Procedures

Each of the 313 participating classroom teachers was asked to provide a copy of his/her most recently administered teacher-made test for a subject other than spelling or math (unless they were teaching secondary mathematics classes). This procedure

resulted in the collection of 175 (56%) usable teacher-made tests which had been recently used in a regular classroom setting.

The sample of 175 teacher-made tests included a total of 6529 test items and 455 item exercises. The cognitive functioning levels of the test items were classified independently by two judges using Bloom's taxonomy of six cognitive demand levels (knowledge, comprehension, application, analysis, synthesis, and evaluation). If these judges differed in their classification of an item or exercise, the item or exercise was reexamined until a consensus was reached.

In addition, each test and each test exercise was examined for format and item construction errors. A test exercise was defined for this study as a group of items of a similar item type on a test. Item construction error criteria were selected from a review of several test construction textbooks designed for preservice education courses. A total of eight item type classifications (completion, essay, multiple-choice, etc.), 10 item format construction error types (does the test have complete directions? are similar type items grouped together? are the items numbered consecutively? etc.), and 66 item construction error types (incomplete stems, implausible alternates, specific determiners, etc.) were derived from these

procedures and were used in the analysis of the teacher-made tests.

An item construction error, if present, was recorded once per item exercise rather than for each time that particular error type may have occurred within the item exercise. In other words, regardless of whether a specific construction error appeared on one item or on several items within the same item exercise a tally of '1' was recorded for that particular error in order to provide a stable base of comparison across tests which varied in number of test items. The tallying of test format errors similarly adhered to the procedure of recording a single tally for each type of format error found on an entire test.

Assessment Instrument

The survey instrument consisted of a 17-item listing of competencies or skills related to teachers' planning and constructing their own tests. These items were constructed by the researchers and then reviewed for appropriateness by a team of five professors responsible for the instruction of the tests and measurements course for preservice teachers at Bowling Green State University. The competency or skills items as they were stated on the assessment form are reproduced in Table 1.

The principals and supervisors were asked to respond via a five-point ('5' as high and '1' as low) Likert-type response

scale for each competency item. The stem for each item response for the principals and supervisors was: "average proficiency of your beginning teachers in this competency." The classroom teachers were asked to denote by the same scale for each of the 17 items: "an estimate of your classroom proficiency in this area." Each respondent was also asked to indicate the nature of his/her school assignment (rural, urban or suburban) and the grade level of his/her assignment (elementary, middle grades, secondary, K-12 grades, or other). Those respondents placing themselves in the "other" and "K-12" categories were excluded from the analyses when the data was examined by grade level assignments of the teachers.

Additionally, the principals and supervisors were asked to comparatively rate the competency level of their beginning teachers in tests and evaluation competencies or skills on three Likert-type five-point scale items ('1' much below average to '5' well above average) relative to: a) subject area knowledge of the teachers, b) teachers' knowledge and skill in other professional education competencies (e.g. planning, discipline, etc.), and c) teachers' overall competencies or skills as educators. The teachers involved in the study were also asked to compare their current tests and measurements skill levels relative to these three areas.

Data Analysis Procedures

The teachers, principals, and supervisors rater classifications were used as the column variable (independent variable) in one-way ANOVA procedures with each of the 17 test construction or planning item "scores" serving as a dependent variable. After a significant F test, Scheffe post-hoc tests ($p \leq .10$) were used to determine any pair-wise mean differences. These procedures were completed separately for those respondents classifying themselves as having primarily an elementary grade level assignment and for those respondents having primarily a secondary grade level assignment.

Independent t-ratio comparisons were used to compare principal and supervisor mean ratings of teachers' proficiency on each of the 17 test planning and test construction competencies or skills and on each of the three comparative proficiency questions. Additionally, Spearman Rho coefficients of correlation were calculated to determine the extent to which the rank ordering of the teachers', principals', and supervisors' competency or skill item rating means were related to one another. Spearman Rho correlations were also calculated between the three sets of perceptual ratings and the data derived from the analysis of actual teacher-made tests.

Results

The average perceptions of the level of teachers' testing proficiencies did not differ significantly when the teacher, principal, and supervisor raters were grouped by school type (rural, urban, and suburban) and by grade level responsibility (elementary and secondary). Further, teachers' mean self-ratings of their current testing proficiencies did not differ when they were grouped by years of teaching experience (1-3, 4-6, and 7-10 years). Similarly, neither the years of teaching experience nor the school type classifications revealed significant differences in the frequency of test construction errors or in the use of the various item type exercises on the sample of teacher-made tests was analyzed. Conversely, however, the grade level analyses revealed significant differences for both the perceptual assessments of the teachers' testing proficiencies and the actual teachers' use of various test item types as displayed on their teacher-made tests. In light of these findings, the results of the analyses of the various ratings and direct assessments of teacher test construction and planning skills are presented separately for grade level responsibility (elementary and secondary) but no further reference will be made to the school type or the years of teaching experience classifications.

Hypothesis One

The series of t-ratio comparisons of principals' and supervisors' ratings of the adequacy of beginning teachers' test construction and test planning skills revealed significant mean differences ($p \leq .05$) on 12 of the 17 competency items and on the combined 17 items ($t=3.34$, $p=.001$) as reported in Table 1. For each of these items revealing a significant difference, the principals rated teachers' testing proficiencies higher than did the supervisors. Similarly, the one-way ANOVA comparisons of teachers', principals', and supervisors' ratings of beginning teachers' proficiencies in test construction and test planning competencies revealed significant differences ($p \leq .05$) among the three groups on each of the 17 competencies for both elementary and secondary teachers. In all instances the teachers' rating mean for each testing competency was significantly higher than the mean of one or both of the administrators' rating means (see Table 2). (The administrators were rating beginning teachers' testing proficiencies while the teachers were rating their current testing proviciencies; however, the teachers' ratings did not differ when classified by years of teaching experience. Thus the writers felt that the comparison of the administrators' and teachers' ratings was meaningful.) These t-ratio and F-test findings of significant

differences between the principals', supervisors', and teachers' ratings resulted in the rejection of hypothesis number one.

In summation the analyses related to the first hypothesis revealed consistent mean differences among the three groups in their ratings of the level of teachers' testing skill proficiencies. The rather consistent pattern of differences indicated that the highest ratings of teachers' testing proficiencies were provided by teachers themselves, followed by principals, and with supervisors providing the lowest proficiency ratings. Although the principals consistently rated the teachers' testing proficiencies higher than did the supervisors, these two administrator groups' ratings were in high agreement regarding the relative skill or proficiency level of the teachers among the various 17 testing competencies. This is clearly evident by a Spearman Rho coefficient of correlation of .92 between the two sets of administrators' ratings (each set of rating means was converted to a rank order and then correlated). Relatedly, the Rho correlation of .68 between the combined ratings of the administrators (principals' and supervisors' ratings together) and the teachers' ratings of their testing proficiencies also suggests a considerable agreement between administrators and teachers about the relative proficiencies of teachers' testing skills.

Hypothesis Two

Each of the 455 test exercises identified on the sample of 175 teacher-made tests was examined for the presence of commonly identified test construction errors. As indicated previously, when an error of a particular type was identified within a test exercise, a single tally was made for that error type regardless of how many specific errors of that type may have occurred within that single exercise. This procedure allowed the investigators to calculate error rates for test exercise item types independent of varying item numbers across tests and exercises.

When the frequency of item construction errors found on the 455 test exercises were tallied for the sample of 175 teacher-made tests, the matching exercises were found to be the most error prone with an average of 6.4 different types of construction errors found on each exercise. Completion exercises were second highest in average number of different types of errors per exercise at 2.2 followed by the essay exercises with an average error rate of 1.5. In descending order of mean frequency error rate, the following exercises were relatively free of construction errors (an average error rate of 1.0 or less): true-false, multiple-choice, short response, problem, and interpretive exercises. The average error rates for the various item type exercises as well as for test format

errors are presented on Table 3, and the types of errors and the frequencies of the occurrence of each error for each item type are reported on Table 4.

To test hypothesis two, the five item types and the test format scale item appearing on the perceptual assessment instrument and which also could be examined on the sample of teacher-made tests were converted to a rank order based upon identified mean error rate (from teacher tests) or perceived proficiency rating mean (from the ratings obtained via the assessment instrument). From direct analyses of the teacher-made tests, the most error free item exercise (multiple-choice) was assigned a rank order of 1 followed by a rank of 2 for the true-false items; rank 3 for the essay items; rank 4 for test format; rank 5 for completion items; and a rank of 6 for the error-prone matching exercises. These ranks were then compared with the ranks derived from the rating means for the principals', teachers', and supervisors' ratings of teachers' testing skills. These rating mean values and associated rank orders for the teachers' testing skills are reported on Table 5, and the associated Spearman Rho correlation coefficients for all pairs of ranks are reported on Table 6.

The Spearman Rho coefficients between the three rater groups were positive with the teachers and supervisors' coefficient highest at .94, followed by the supervisors and

principals' coefficient at .64, and by the teachers and principals' coefficient at .41. The Spearman Rho coefficients between each of the three groups of rank ordered rating means and the rank order of mean error frequencies found on the actual teacher-made tests resulted in three negative coefficients. This indicates a negative relationship between the perceptual quality ratings that the raters gave teachers and the frequencies of test item type and test format errors found on the actual teacher-made tests. The rank ordering of teacher test construction and format proficiencies derived from the direct assessment of teachers' tests correlated -.71 with supervisors' ratings, -.60 with teachers' ratings, and -.50 with principals' ratings of these same teacher testing proficiencies.

Illustrative of the differences between the perceptual ratings of teachers' testing skills and the errors found from the direct assessment of teachers' testing skills as displayed on their tests are: the teachers', principals', and supervisors' each rated teachers' proficiencies in writing matching test exercises higher than teachers' proficiencies in writing the four other item types (multiple-choice, completion, true-false, and essay); whereas the direct analysis of the teacher-made tests revealed that the matching exercises contained more than twice as many errors as compared to each of the four other item types. Similarly, the three rater groups

indicated that teachers' test construction proficiencies were relatively low for the writing of multiple-choice and true-false item types; whereas the direct assessment of the teachers' tests indicated that on the average these item types were relatively free of test construction errors with an average of one or fewer errors per exercise.

Relative to writing test items which function at higher cognitive levels, the principals and supervisors rated teachers' proficiency in this skill lowest relative to the 17 testing competencies assessed, the elementary teachers rated this proficiency among their lowest testing skills (15th of 17), but the secondary teachers rated this writing proficiency among the top 25% of their 17 testing skills. The direct analyses of the teacher-made tests supported the low ratings for this competency. With the exception of the math and science tests, approximately 90% of the teacher-made tests consisted of items functioning almost exclusively at the knowledge level (90-100% demanding just simple recall). As most of the test items functioning at higher cognitive levels were found on the secondary math and science tests, this may in part account for the secondary teachers' higher ratings of their proficiency in writing items demanding higher cognition responses as compared to the two administrator and the elementary teacher rater groups.

In summation, the lack of agreement between the perceptual ratings (rankings) of the teachers, principals, and supervisors as compared to the rankings derived from the frequencies of actual errors observed on the teacher-made tests led to the rejection of hypothesis number two. The moderately high negative correlations found between the perceptual quality ratings of teachers' testing skills by the teachers, principals and supervisors and the results of the direct assessments of the teacher-made tests indicate a marked discrepancy between perceived quality of teachers' testing skills and teachers' actual test item and test format writing proficiencies as displayed on their tests. There was higher agreement, however, between the perceptual ratings and observed proficiencies of teachers' abilities to write test items functioning at higher cognitive levels even though elementary and secondary teachers differed one from the other in rating this skill.

Hypothesis Three

The teachers', principals', and supervisors' ratings of teachers' testing and evaluation skills when compared to ratings of teachers' knowledge of their subject areas, their professional education competencies or skills, and their overall competence as educators revealed that all three rater groups perceived beginning teachers to be less proficient in testing skills than in these three other areas of professional

competence. The teachers', principals', and supervisors' rating means for this section of the assessment instrument are reported on Table 7.

The responses to this section of the questionnaire were analyzed by the raters' grade level responsibility, by raters' type of school assignment, and by raters' administrative position (principal and supervisor). When the total group of respondents was classified by grade assignment (elementary, middle, or secondary schools) and by type of school (rural, urban, or suburban), no significant rating mean differences were noted for the three rating items. However, the principals' and supervisors' rating means were found to differ significantly for each of the three items. The item rating means for each of the three items for the principals' and supervisors' ratings were: teachers' knowledge of their subject area, principals 3.03, supervisors 2.87 ($t=2.47$, $p=.02$); teachers' other professional education competencies or skills, principals 2.96, supervisors 2.81 ($t=2.34$, $p=.02$); and teachers' overall competencies as educators, principals 2.93 and supervisors 2.73 ($t=3.34$, $p=.001$). On each of these three items the supervisors' mean rating of the beginning teachers' competencies was significantly lower than the principals' mean rating. The teachers' mean rating for these three items, 2.86, 2.91 and 2.80 respectively, did not differ significantly from the administrators' rating

means on these three items as they did on the 17 teachers' test competency items. The principals' rating means though were higher than that of the supervisors as they were for 12 of the 17 teacher testing competency items.

In summation, the third hypothesis was rejected as the item rating means for the principal, supervisor, and the teacher respondents were below average (below 3.0) for eight of the nine group rating means. The principals, supervisors, and teachers perceived beginning teachers (or their current proficiencies in the case of the teachers) as being somewhat less proficient in testing and evaluation skills as compared to teachers' knowledge and skills in the other three listed professional competency areas.

Summary, Implications, and Discussion

The analyses of the data collected from the teachers, principals, supervisors and from the direct assessments of the teacher-made tests resulted in the rejection of each of the three stated hypotheses. It was found that teachers, principals, and supervisors do not agree in their assessments of the levels of teachers' test construction and test planning proficiencies. The classroom teachers rated their testing proficiencies significantly higher than did the administrators, and the principals rated teachers' testing proficiencies significantly higher than did the supervisors. Secondly, the

data gathered revealed that there were marked discrepancies (negative correlations) between perceptual quality ratings of teachers' testing skills by principals, supervisors, and teachers and the direct assessments of teachers' testing proficiencies as displayed on their tests. And third, the data gathered revealed that principals, teachers, and supervisors perceived beginning teachers' testing and evaluation proficiencies (or current proficiencies in the case of teachers) to be somewhat below average when compared to teachers' subject area knowledge, teachers' other professional education competencies and skills (planning, discipline, etc.), and teachers' overall competence as educators.

The findings from this investigation appear to have several possible implications for both those educators concerned with teacher preservice and those concerned with the inservice training of teachers:

1. The results of this investigation would suggest that principals', supervisors', and teachers' perceptual ratings of teachers' test-making proficiencies may not be very accurate guides for determining teachers' inservice training needs as compared to direct assessments of these test-making proficiencies as displayed on teacher-made tests.

2. Data obtained from this investigation would suggest that principals', supervisors', and teachers' perceptual ratings of teacher test-making proficiencies differ in magnitude but show a rather high relative relationship with each other. This suggests that inservice planners should focus on relative differences among skill proficiencies within the rater groups in determining teacher inservice needs rather than comparing or combining rating magnitudes for a particular skill across rater groups.
3. Preservice and inservice trainers are probably safe to assume that teachers' test planning skills are at about the same level of proficiency as teachers' ability to write test items and to develop test formats free from error. The principals, supervisors, and teachers in this study rated teachers' test planning skills at about the same level as teachers' test item writing skills. As these groups appeared to overestimate teachers' actual testing skills as displayed on their tests, it is likely teachers' actual test planning skills may also be at a much lower proficiency level than perceived by principals, supervisors, and teachers.
4. The results of this study would suggest to those planning teacher inservice training curricula and activities that teacher test construction skills do not vary significantly

by school type (rural, urban, and suburban), teachers' years of teaching experience (at least from 1 to 10 years), and by grade level assignment of teachers although teachers' preferences for items or at least use of item types was found to vary significantly by grade level and subject area of specialization. (Most teachers, however, reported using a variety of item types in constructing their tests.) Thus, basic preservice and inservice test construction curricula are likely to be equally effective in a variety of school settings for a variety of teachers.

5. Both preservice and inservice teacher trainers need to increase attention being given to item writing skills and to the writing of items that measure beyond simple knowledge levels. The findings from this investigation suggest that neither beginning teachers nor teachers with up to 10 years of teaching experience display high levels of proficiency in these skills on their tests. In this regard, the data from this study appeared to generally support the findings reported by Billeh (1974) that the more experienced as compared to less experienced teachers (this difference in this study only approached significance) in fact write tests with proportionately more knowledge than items at higher cognitive levels and the findings of Fleming and Chambers (1983) that high cognitive

functioning items are found almost exclusively on math and science tests.

6. Preservice and inservice teacher trainers in particular need to emphasize the careful construction of matching exercises. The results of this study support the reputation of matching exercises in the measurement literature as being the most error prone item type (Gronlund, 1985). The matching exercises in the sample of tests collected for this investigation were found to have two or more times as many errors per exercise as compared to the other item types examined.
7. Many of the test item and test format construction errors found on the teacher-made tests were nontechnical in nature such as lack of directions, lack of consecutive item numbering, insufficient margins or spacing between items, illegible handwriting, etc. as was also reported by Fleming and Chambers (1983). This would suggest that many types of errors on teachers' tests might be readily addressed by school personnel and other inservice trainers who do not have extensive expertise in test construction.
8. The relatively low principals', teachers', and supervisors' ratings of teachers' test construction and test planning proficiencies relative to their other professional competencies would suggest that more preservice and

inservice instruction is needed in this area. Further, as these three groups of inservice educators have an acknowledged awareness of lower teachers' proficiencies in testing skills as compared to other skills, one would logically presume that each of these groups of educators would be supportive of teacher inservice training designed to develop teachers' testing skills. And as a further thought, the lack of increased proficiency levels in testing skills reported by teachers with from one to ten years of teaching experience may suggest either that little inservice training related to test construction is being given or that such training has not been successful.

The readers are cautioned that the generalization of the results of this study may be limited by a sample of teachers, principals, and supervisors from a single state, by a sample of teachers and their teacher-made tests obtained from graduates from a single institution, by an approximate 65% response rate from the subjects selected for study, and by principals' and supervisors' ratings of their typical beginning teachers rather than of their total teaching faculty. Additionally and because of the nature of the survey assessment instrument used in this investigation, only a limited number of comparisons were made between the teachers' test construction skills as rated by the principals, supervisors, and teachers and those teachers' test

construction skills as assessed by the direct analyses of the teacher-made tests. Conversely however, a reader might conclude that some basis for generalization of these findings to a larger population exists because the findings from this investigation appear to be consistent with those of other studies, neither the teachers' self-ratings nor their skills as displayed on their teacher-made tests varied markedly when classified from 1 to 3, 4 to 6, and 7 to 10 years of teaching experience, the teachers having graduated from the single institution had graduated from that institution over a full decade during which twenty or more different professors had shared instructional responsibilities for the required preservice tests and measurements course, and as relatively large samples of educators employed in a diverse variety of school settings served as subjects for this study.

8/3

References

Black, T. R. (1980). An analysis of levels of thinking in Nigerian science teachers' examinations. Journal of Research in Science Teaching, 17, 301-306.

Billeh, V. Y. (1974). An analysis of teacher-made science test items in light of the taxonomic objectives of education. Science Education, 58, 313-319.

Dwyer, C. A. (1982). Achievement testing. In H. E. Mitzel (Ed.), Encyclopedia of Educational Research (4th ed., Vol. 1, pp. 13-22). New York: The Free Press.

Fleming, M., & Chambers, B. (1983). Teacher-made tests: Windows on the classroom. New Directions for Testing and Measurement, 19, 29-38.

Gronlund, N. E. (1985). Measurement and Evaluation in Teaching (5th ed.). New York: MacMillan Publishing Company.

Gullickson, A. R. (1984). Teacher perspectives of their instructional use of tests. Journal of Educational Research, 77, 244-248.

Gullickson, A. R. (1986). Teacher education and teacher-perceived needs in educational measurement and evaluation. Journal of Educational Measurement, 23, 347-354.

Gullickson, A. R., & Ellwein, M. C. (1985). Post hoc analysis of teacher-made tests: The goodness-of-fit between prescription and practice. Educational Measurement: Issues and Practice, Spring, 15-18.

Lambert, R. F. (1980-81). Teacher attitudes on testing: A multiple perspective. College Board Review, 29-30, 13-14.

Rogers, B. G. (1985). Prospective teacher perceptions of how classroom teachers use evaluation methods: A qualitative research approach. Mid-western Educational Researcher, 613-620.

Salmon-Cox, L. (1981). Teachers and standardized achievement tests: What's really happening. Phi Delta Kappan, 62, 631-634.

Stiggins, R. J., & Bridgeford, N. J. (1985). The ecology of classroom assessment. Journal of Educational Measurement, 22, 271-286.

8/3

Table 1

Comparisons of Principals' and Supervisors' Ratings of Teachers' Test Construction and Planning Proficiencies*

Competencies or Skills	Rating Means			
	Prin.	Super.	t	p
1. Writing good multiple-choice questions	3.06	2.91	2.67	.008
2. Writing good completion questions	3.13	2.97	2.60	.010
3. Writing good matching questions	3.16	3.04	2.07	.039
4. Writing good true-false questions	3.06	2.90	2.50	.013
5. Writing good essay questions	2.85	2.59	3.69	.001
6. Scoring essay questions	2.78	2.53	3.42	.001
7. Identifying good and poor questions for future tests	2.92	2.73	2.98	.003
8. Writing questions in harmony with school and class goals	2.88	2.72	2.19	.029
9. Stating objectives sufficiently clear to suggest test items	2.97	2.73	3.34	.001
10. Writing test questions that demand higher thinking processes	2.65	2.43	2.91	.004
11. Constructing tests that represent true student progress	2.88	2.65	3.27	.001
12. Use of less formal assessments: checklists, ratings, etc.	2.93	2.79	2.15	.032
13. Use of observations (visual) to assess and guide learning	3.02	2.91	1.62	.106
14. Use of sociometric, guess who, and related techniques	2.73	2.76	0.41	.680
15. Selecting good test questions from teacher manuals	3.16	3.12	0.66	.511
16. Setting up readable, scorable, and attractive tests	3.05	3.01	0.64	.523
17. Making tests reflect what is covered in text and class	3.24	3.14	1.53	.127
	50.48	47.88	3.34	.001

*All principals' responses, regardless of grade level responsibility, were combined for these analyses; likewise for the supervisors.

Table 2

Comparisons of Secondary and Elementary Principals', Supervisors', and Teachers' Ratings of Teachers' Test Construction and Planning Proficiencies*

	Secondary Means						Elementary Means					
	Prin. (1)	Supr. (2)	Tchr. (3)	F	p	Sch.**	Prin. (1)	Supr. (2)	Tchr. (3)	F	p	Sch.**
1.	3.06	2.95	3.71	31.35	.001	3>1,2	3.07	2.91	3.56	15.69	.001	3>1,2
2.	3.12	2.92	3.84	34.15	.001	3>1,2	3.15	3.03	3.53	7.87	.001	3>1,2
3.	3.15	3.02	3.92	44.47	.001	3>1,2	3.16	2.92	3.62	13.38	.001	3>1,2
4.	3.01	2.84	3.56	20.99	.001	3>1,2	3.11	2.99	3.49	5.78	.01	3>1,2
5.	2.87	2.47	3.67	36.18	.001	3>1,2	2.86	2.74	3.16	3.65	.05	3>2
6.	2.83	2.45	3.45	23.47	.001	3>1,2	2.76	2.55	2.84	2.07	.13	---
7.	2.90	2.67	3.85	47.59	.001	3>1,2	2.98	2.78	3.51	13.54	.001	3>1,2
8.	2.84	2.71	3.78	36.19	.001	3>1,2	2.93	2.76	3.57	14.90	.001	3>1,2
9.	2.95	2.85	3.63	19.39	.001	3>1,2	3.01	2.58	3.40	14.59	.001	3>1,2
10.	2.67	2.44	3.86	61.03	.001	3>1,2	2.67	2.47	3.27	11.48	.001	3>1,2
11.	2.81	2.56	3.68	38.83	.001	3>1,2	2.98	2.73	3.43	11.48	.001	3>1,2
12.	2.90	2.72	3.13	4.67	.02	3>2	2.97	2.85	3.27	4.16	.05	3>1,2
13.	3.00	2.90	3.44	9.68	.001	3>1,2	3.05	2.99	3.67	12.28	.001	3>1,2
14.	2.75	2.75	3.13	5.92	.004	3>1,2	2.74	2.78	3.27	8.90	.001	3>1,2
15.	3.09	3.16	3.80	23.44	.001	3>1,2	3.25	3.19	3.51	3.36	.05	3>1,2
16.	3.12	2.97	4.03	46.41	.001	3>1,2	3.07	3.00	3.60	9.01	.001	3>1,2
17.	3.29	3.18	4.35	58.36	.001	3>1,2	3.24	3.18	3.76	8.21	.001	3>1,2
++	50.36	47.31	63.14	71.39	.001	3 1,2	51.06	48.27	58.15	14.30	.001	3>1,2

* See Table 1 for description of competencies 1-17

** Results of the Scheffe post-hoc tests at $p \leq .10$.

++ Totals all items combined

Table 3

Mean Frequency of Test Item or Test Format Construction Errors Per Item Type Exercise or Test

	No. Items Reviewed	% Total Items Reviewed	No. of Exercises	No. Errors Present*	Mean Errors Per Exercise
Item Type Errors					
1. Matching	1261	19	78	496	6.4
2. Completion	549	8	48	106	2.2
3. Essay	64	1	22	34	1.5
4. True/False	935	14	69	71	1.0
5. Multiple-Choice	1317	20	65	53	0.8
6. Short Response	1093	17	89	61	0.7
7. Problems	896	14	54	26	0.5
8. Interpretive Exercise	362	6	30	6	0.2
9. Unclassified	52	1	6	--	--
Subtotals	6529	99	455	853	1.9
10. Test Format Errors (175 test formats reviewed)				281**	1.6++

*Each specific item type construction error was tallied only once if present in an exercise (i.e., an error may have occurred several times or once in an exercise but in either case only a single tally was used so that tests and exercises could be compared regardless of the number of individual items appearing in a test or exercise).

**There were only 175 individual tests but some tests had more than one format error.

++Mean frequency of format errors per test.

Table 4

Frequency and Nature of Item Construction Errors Found on Each Item Exercise

Construction Error	N	%*
a. Completion Item Type		
Not complete interrogative sentence	32	30
Blanks in statements	31	29
Textbook statements with words left out	18	17
More than single blank in statement	12	11
Question allows more than single answer	6	6
Blank number clue	4	1
Blank length clue	1	1
Requests trivia versus significant idea	1	1
Unstated degree of precision	1	1
Lengthy, unnecessary words or phrases	0	0
	106	100
b. True-False		
Required to write response, time waste	20	28
Statements contain more than single idea	16	23
Negative statements used	15	21
Presence of specific determiner	8	11
Statement not question, give away item	6	8
Needless phrases present, too lengthy	4	6
Imprecise statement, not always true or false	1	2
Presence of length clue	1	1
Opinion not attributed to source	0	0
	71	100
c. Essay Exercises		
Response expectations unclear, not labeled, etc.	14	41
Scoring points not realistically limited	7	21
Optional questions provided	5	15
Restricted question not provided	3	9
Ambiguous words used	2	6
Opinion or feelings requested	2	6
Question limited to simple listing response	1	2
	34	100
d. Problem Exercises		
Items not sample understanding concepts, only calculations	20	77
Not range of easy to difficult problems	3	12
Degree of accuracy not requested	2	8
Nonindependent items	1	4
Use of objective items when calculation preferable	0	0
	26	100

(table continues)

Construction Error	N	%*
e. Matching Item Type		
Columns not titled	71	14
Not use one, more than once, or not all not in directions to prevent elimination	69	14
Response column not ordered	60	12
Directions not specify basis for match	55	11
Answering procedure not specified	52	10
Elimination due to equal numbers	46	9
Column(s) exceed 10 items	39	8
Materials not homogeneous	38	8
Premise not to left side	37	7
Numbers not to left and letters to right	13	3
Exercise not contained on single page	7	2
Requires responses to be written out	6	1
Insufficient information in premises	3	1
	496	100
g. Interpretive Exercises		
Objective response form not used	6	100
Can be answered without data presented	0	0
Errors present in response items	0	0
Data presented unclear	0	0
	6	100

Construction Error	N	%*
f. Multiple Choice		
Alternates not in column(s) or rows	21	40
Incomplete stems	12	23
Negative words not emphasized or avoided	9	17
"All or none above" not appropriately used	5	9
Needless repetition in alternates	2	4
Presence of specific determiners in alternates	2	4
Verbal associations between alternate and stem	1	1
Alternates overlap	1	1
Needless phrases used	0	0
Grammatical clues	0	0
Distractors implausible	0	0
Length clues	0	0
a and c, but not b, etc. used	0	0
	53	100
h. Short Response		
Item requires only listing	51	84
Response expectations ambiguous, not specified	7	11
Unrealistically high scoring values assigned	3	5
	61	100

*Each specific item type construction error was tallied only once if present in an exercise (i.e., an error may have occurred several times or once in an exercise but in either case only a single tally was used so that tests and exercises could be compared regardless of the number of individual items appearing in a test or exercise), the percentage refers to percent of this error type to all errors found on all exercises of this type.

Table 5

Two Sets of Rank Orders of Teachers' Testing Competencies: (1) Based on Teachers', Principals', and Supervisors' Perceptual Ratings and (2) Based Upon Analysis of Actual Teacher-Made Tests

	Perceptual Rating Means						Analysis Actual Tests	
	Teachers		Principals		Supervisors			
Competencies	Mean	Rank	Mean	Rank	Mean	Rank	Mean*	Rank
1. Writing multiple-choice items	3.64	4	3.06	3.5	2.91	4	.8	1
2. Writing completion items	3.72	3	3.13	2	2.97	3	2.2	5
3. Writing matching items	3.81	2	3.16	1	3.04	1	6.4	6
4. Writing true-false items	3.58	5	3.06	3.5	2.90	5	1.0	2
5. Writing essay items	3.37	6	2.85	6	2.59	6	1.5	3
6. Test format	3.88	1	3.05	5	3.01	2	1.6	4

*Mean number of test construction errors per exercise (or per test for test format errors) found on the teacher-made tests

Table 6

Spearman Rho Correlations Between Perceptual Ratings Ranked by Mean Magnitudes and Ranked Construction Error Frequencies Observed on Teacher-Made Tests

Source of Ranking	(1)	(2)	(3)	(4)
1. Teachers' Ratings	1.00	.94	.41	-.60
2. Supervisors' Ratings		1.00	.64	-.71
3. Principals' Ratings			1.00	-.50
4. Test Error Frequencies				1.00

Table 7

Beginning Teachers' Testing Proficiencies* Compared to Their Other Proficiencies, As Rated by Principals, Supervisors, and Teachers

Relative Proficiency Rating Items**	Rating Means		
	Principals	Supervisors	Teachers
1. Relative to knowledge of their subject areas, beginning teachers' test and evaluation competencies are...	3.03	2.87	2.86
2. Relative to their other professional education competencies, such as planning, discipline, etc., beginning teachers' test and evaluation competencies are...	2.96	2.81	2.91
3. Relative to their overall competencies as educators, beginning teachers' test and evaluation competencies are...	2.93	2.73	2.80

*Regardless of grade level

**Ratings were recorded via a five point Likert-type scale, 5 (well above average), 4 (somewhat above average), 3 (about average), 2 (somewhat below average), and 1 (much below average)

8/3